Progress Report: Week II

Esperanto

Exchanges for SPEech ReseArch aNd TechnOlogies

Horizon 2020 project

Encoding team

Overview

- Preliminary results on speech translation (21 X -> En) tasks using an encoder/decoder framework
  - ⇒ Encoder initialisation : SAMU-XLS-R vs XLS-R
  - ⇒ Decoder initialisation : randomly initialized vs mbart[1] initialisation
  - ⇒ Encoder fine-tuning strategy : full fine-tuning vs adapters
  - ⇒ Decoder fine-tuning strategy : full fine-tuning (randinit), encoder attention and layer norm fine-tuning (mbart decoder)

[1] Yinhan Liu and Jiatao Gu and al. 2020. Multilingual Denoising Pre-training for Neural Machine Translation [Online] Available: https://arxiv.org/pdf/2001.08210

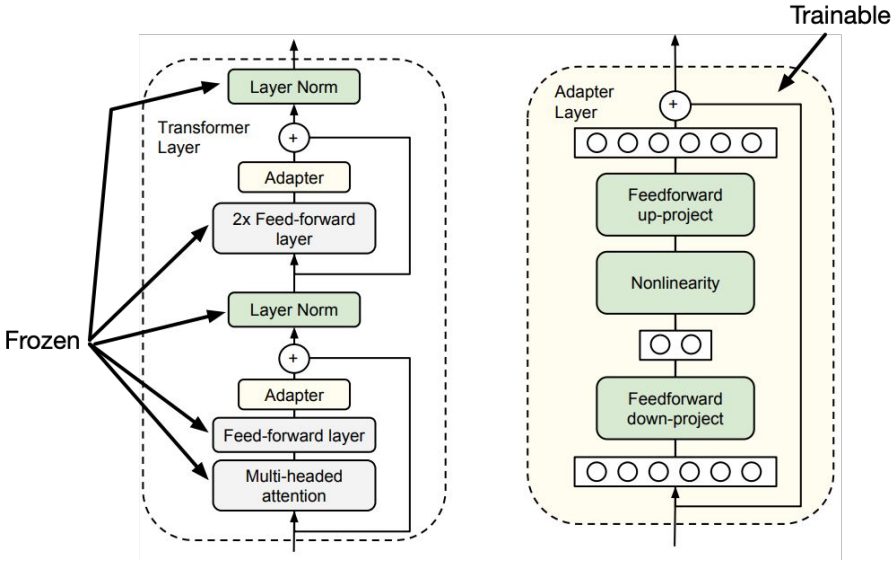
Randomly initialized decoder

- Random initialization of the decoder (6 transformer layers, 16 attention heads, 1024 embedding), fully fine tuning the encoder

⇒ XLS-R (graph LR, MR, HR):

⇒ SAMU-XLSR :

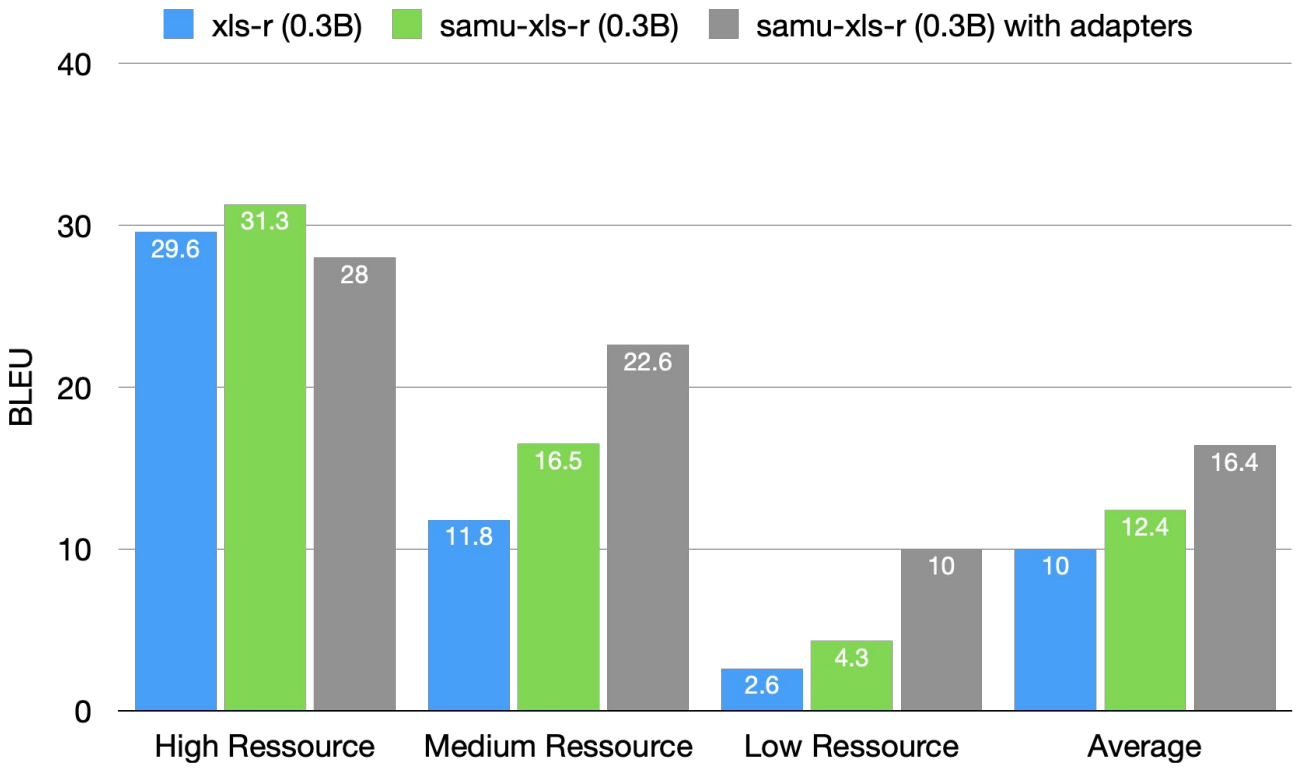
Parameter-efficient fine-tuning with Adapters [1]



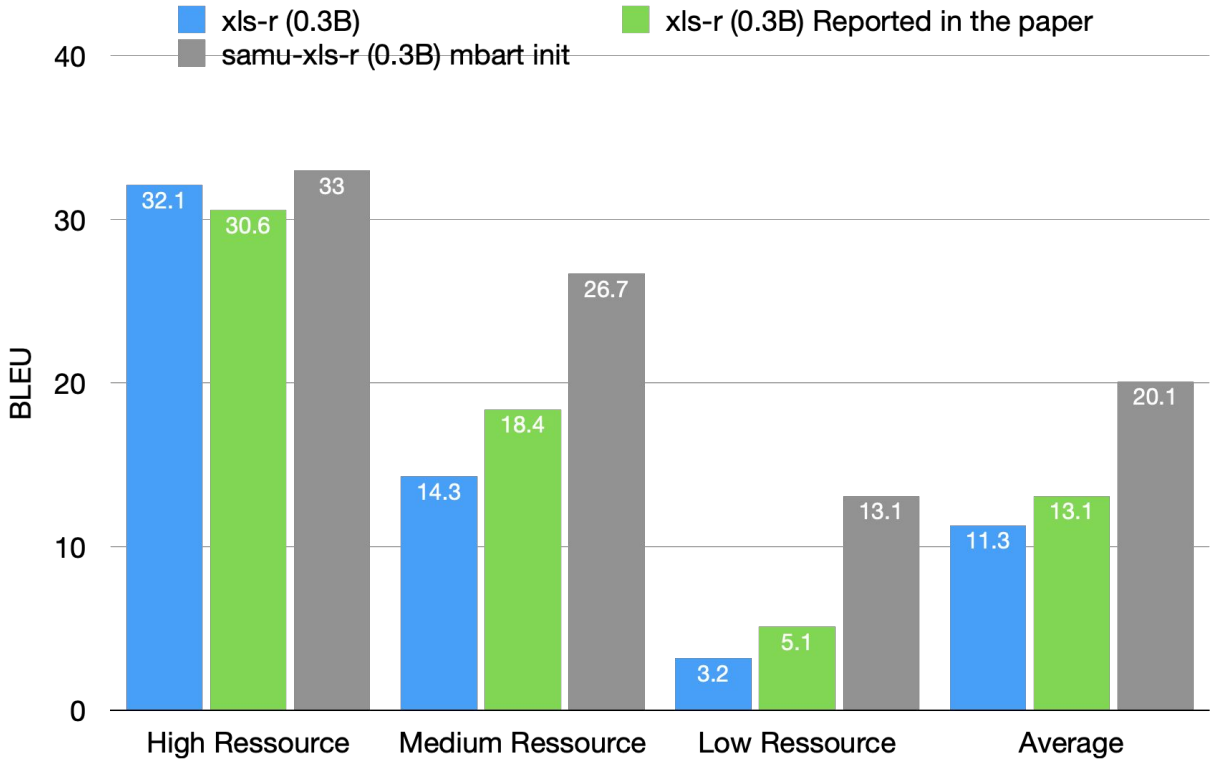
- Total trainable parameters :
  ⇒ 700M
- Trainable parameters with adapter fine-tuning :
  ⇒ 56M

[1] Neil Houlsby, Andrei Giurgiu, et al. Parameter-Efficient Transfer Learning for NLP, ICML 2019

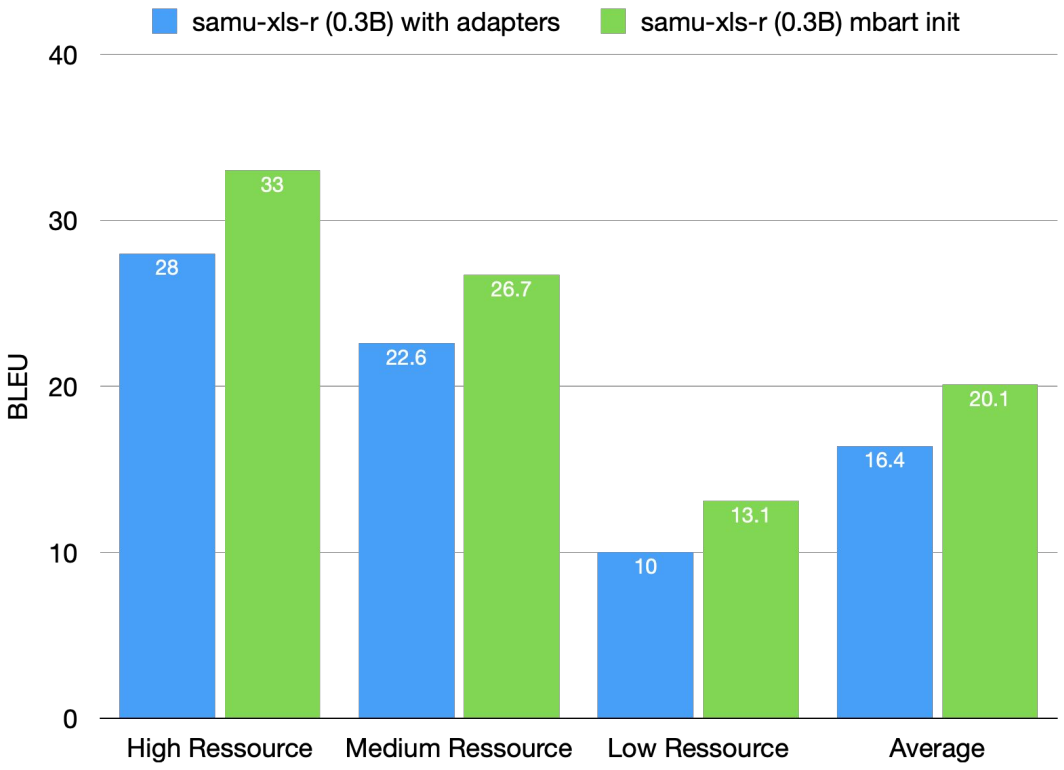
Results with randomly initialized decoder



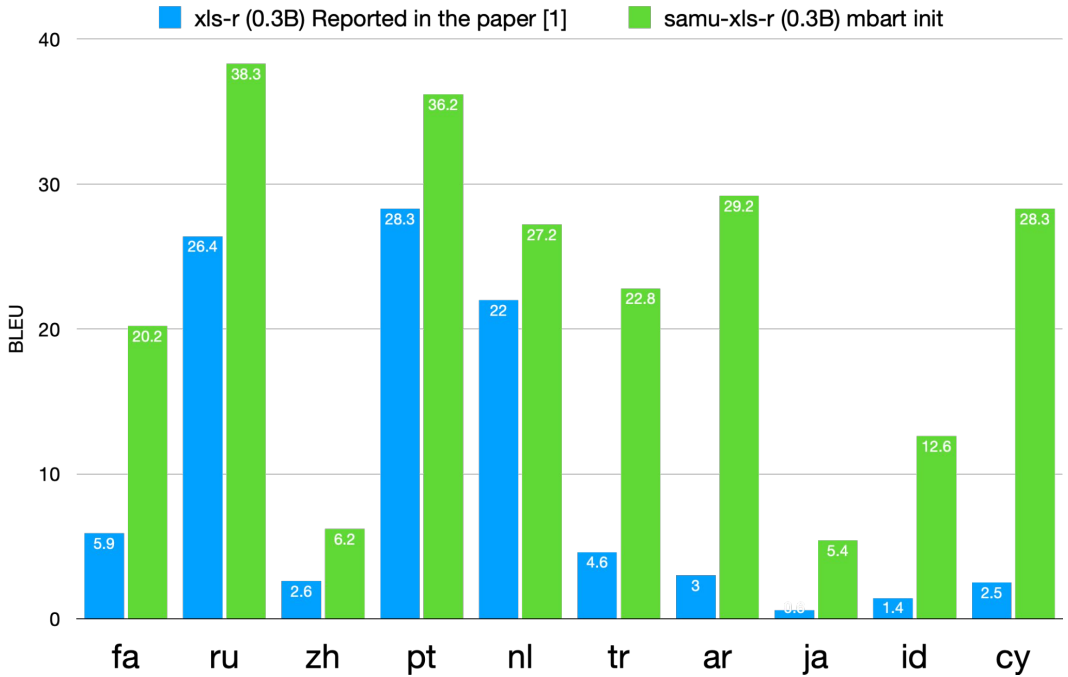
Results with mbart decoder



Mbart decoder vs randomly initialized decoder



XLS-R vs SAMU-XLS-R on individual languages



[1] Babu, A., Wang, C., Tjandra, A et al. (2021). XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

What's next ?

- ASR fine-tuning of XLSR → downstream speech translation task fine-tuning

  ⇒ Leads to fairer comparison with SAMU-XLS-R

- Evaluate SAMU-XLSR embeddings on speech-to-speech retrieval

  ⇒ To better understand low-performance on some X→EN speech translation tasks (Chinese, Japanese, …)

Analysis

Analysis

- **<u>Goal:</u> help the encoder team building a better encoder**
  - identifying how and where the embeddings are lacking cross-modal and cross-lingual features

- **Common space probing tasks:**
  - Layer-wise speaker verification
  - Layer-wise emotion classification
  - Cross-modal/Cross-lingual semantic understanding probing
  - Language identification
- **Direct analysis of the common space:**
  - Difference in embeddings from different models
  - Framewise embeddings alignment

Analysis
Overview

- First results on probing tasks
  - layer-wise emotion recognition
  - layer-wise spoken language understanding
  - cross-lingual spoken language understanding

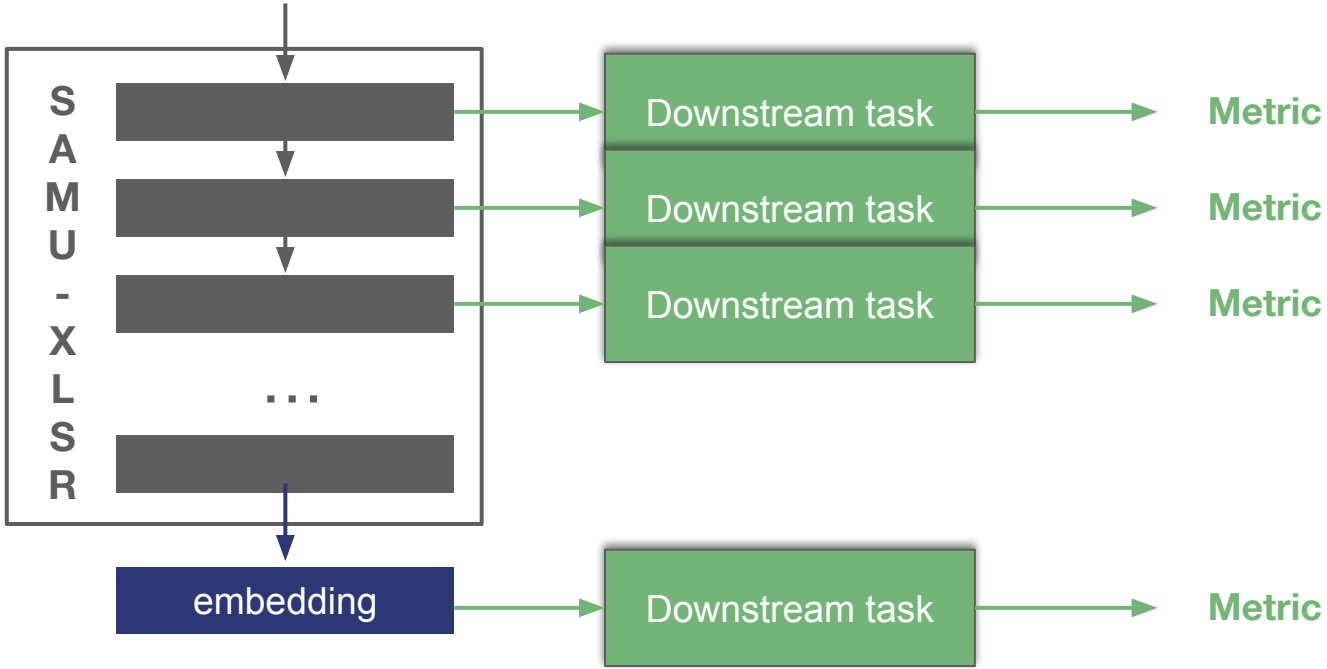
Analysis
Some results

- Probing the presence (or absence) of specific information in the encoder
  - Speaker verification
  - Emotion classification
  - Semantic concepts detection
- At which layer in the encoder are those information present?

- Are the common space semantic information cross-lingual?
  - Training a SLU module on speech French samples, testing on Italian

Analysis

Layer-wise probing task



Analysis
Layer-wise emotion classification

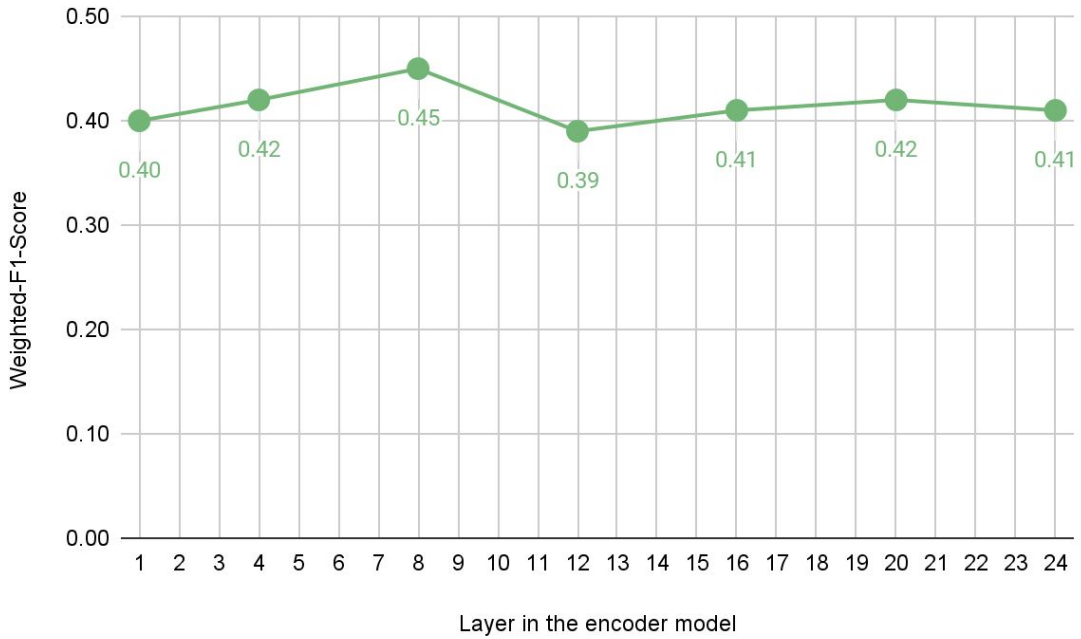
	Accuracy	F1-Score	Weighted-F1-Score
Baseline	*0.343*	*0.176*	*0.34*
Layer 1	0.406	0.221	0.40
Layer 4	0.434	0.250	0.42
Layer 8	0.453	0.234	**0.45**
Layer 12	0.437	0.192	0.39
Layer 16	0.410	**0.254**	0.41
Layer 20	0.439	0.238	0.42
Layer 24	0.440	0.226	0.41

- Input embeddings:
  - XLS-R (freezed)
- Downstream model:
  - ECAPA-TDNN (4 SE-Res2Net)

- Emotion information is present (and kept) in all the layers

Analysis

Layer-wise emotion classification



Input embeddings:
- XLS-R (freezed)

Downstream model:
- ECAPA-TDNN (4 SE-Res2Net)

Emotion information is present (and kept) in all the layers

Analysis

Layer-wise semantic concepts detection (SLU)

- French MEDIA SLU dataset

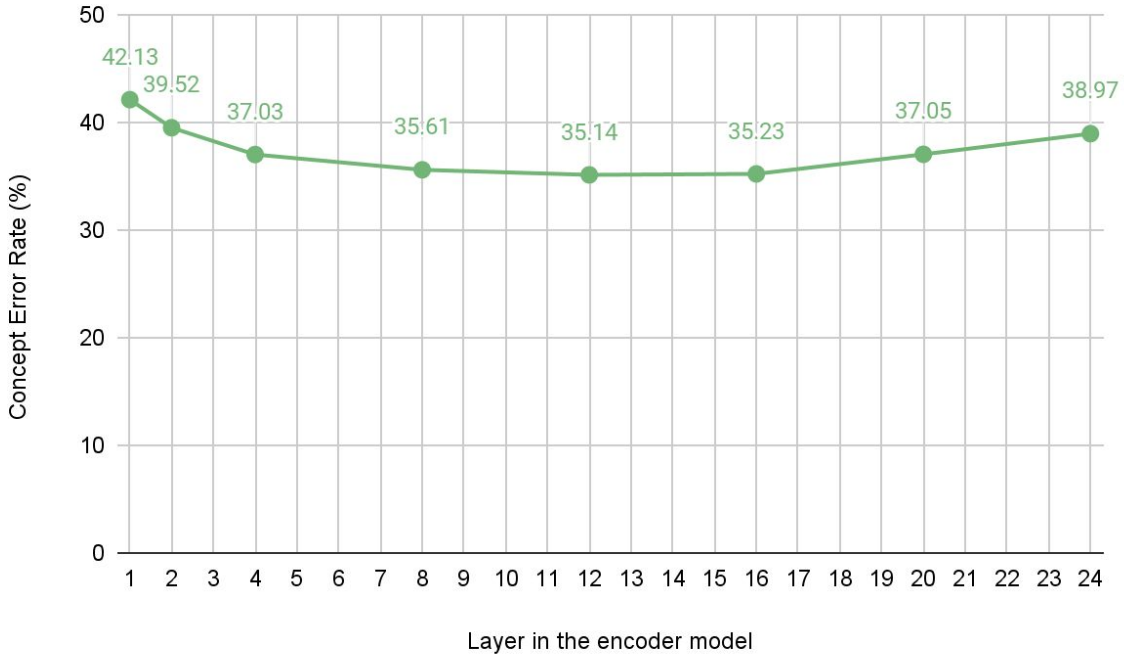
  `I will check-in at <hour-arrival noon >`

	Concept Error Rate
Layer 1	42.13
Layer 2	39.52
Layer 4	37.03
Layer 8	35.61
Layer 12	**35.14**
Layer 16	35.23
Layer 20	37.05
Layer 24	38.97

- Freezed input embeddings (XLS-R)
- 3 Bi-LSTM layers
- 3 linear layers
- CTC loss

Analysis

Layer-wise semantic concepts detection (SLU)



- French MEDIA SLU dataset
  `I will check-in at <hour-arrival noon >`
- Freezed input embeddings (XLS-R)
- 3 Bi-LSTM layers
- 3 linear layers
- CTC loss

Analysis

Cross-lingual probing with a SLU module

- French MEDIA SLU dataset

  `<answer non > merci`

- Italian PortMEDIA SLU dataset
  - same domain
  - but not a parallel alignment with MEDIA

  `<answer sí > per favore`

	Test Concept Error Rate
MEDIA (FR)	27.09%
PortMEDIA (IT)	

- Freezed input embeddings (SAMU-XLSR / LaBSE)
- 3 Bi-LSTM layers
- 3 linear layers
- CTC loss
- Greedy decoder

Analysis

SAMU-XLSR (speech) vs LaBSE (text) embeddings

- SAMU-XLSR:
  - XLS-R embeddings projected into LaBSE embedding space
- Is a model able to discriminate between LaBSE and SAMU-XLSR embeddings? YES

	Embedding discrimination accuracy
In-Domain languages	100%
Out-of-Domain high resource languages	55-60%
Out-of-Domain low resource languages	85-95%

Analysis
Next goals

- How can the encoder can make the transition from
  - the acoustic space (speaker, phonetic, …)
  - to the semantic space ?
  - with either
    - different languages
    - different sentences
    - different speakers
  - At which layer?
- Semantic probing:
  - Use LaBSE to probe for cross-modal information

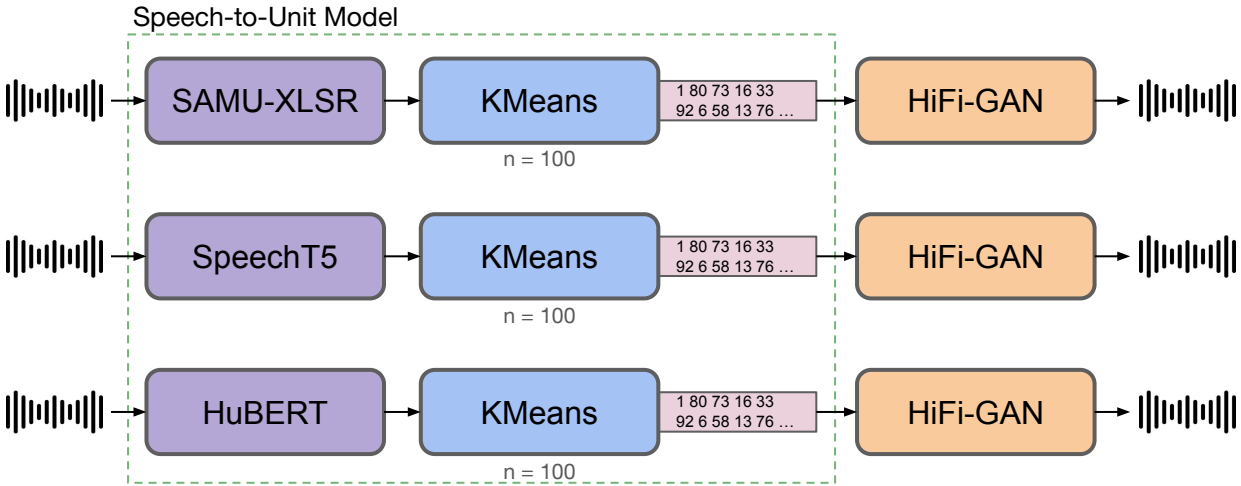
Decoding Team
Overview

- Preliminary results in speech generation
- Preliminary results in text generation
- Decoder architecture work in progress

Decoding Team
Speech Generation

Objective : Generate speech directly from self-supervised discrete representations

- Generating speech from HuBERT representation

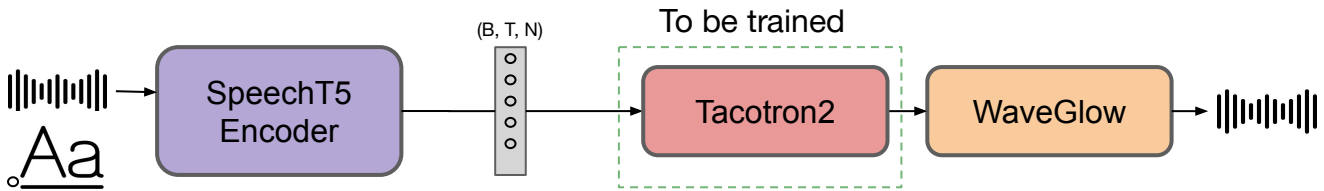


[1] A. Polyak et al., " Speech Resynthesis from Discrete Disentangled Self-Supervised Representations" 2021. [Online]. Available: https://arxiv.org/abs/2104.00355

Decoding Team
Speech Generation

Objective : Compare synthetic speech from SAMU-XLSR/Labse (multilingual embeddings) and SpeechT5 (monolingual embeddings)

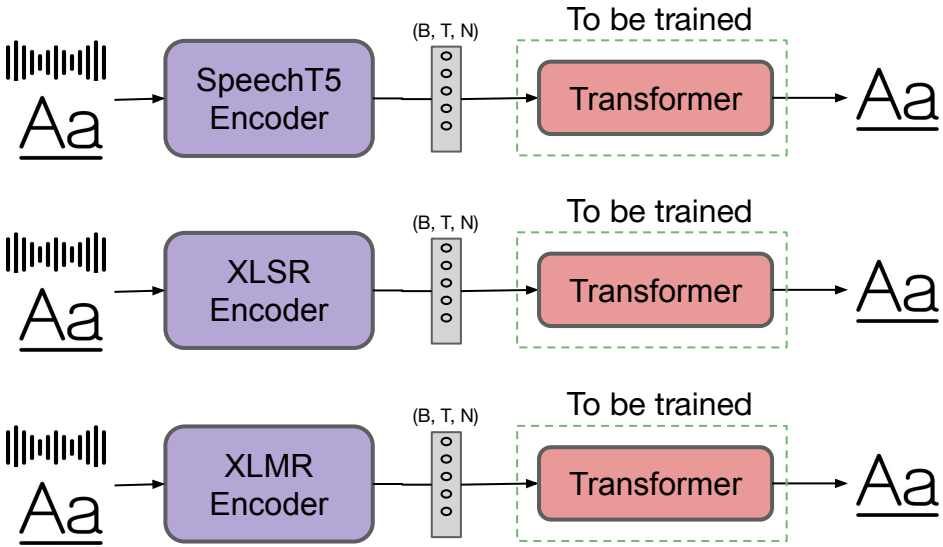
- SpeechT5 using speech-only
- SpeechT5 using text-only input
- SAMU-XLSR
- LaBSE token level embeddings



[1] J. Ao et al., " SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing," 2021. [Online]. Available: https://arxiv.org/abs/2110.07205

Decoding Team
Text Generation

Objective : Compare text generation from embedding from XLSR, XLMR and SpeechT5 (monolingual embeddings)



Decoding Team
Decoder architecture